**Практическая работа 3 «Модели с бинарными зависимыми переменными**

1. **Запрет курения на рабочем месте**

**Исследовательский вопрос**

Ваша задача — исследовать, действительно ли запрет курения на рабочем месте побуждает курильщиков избавляться от этой вредной привычки.

**Набор данных**

В этом задании необходимо оценить влияние запретов на курение на рабочем месте, используя базу данных по 10 тыс. офисным работникам в США в период с 1991 по 1993 год, содержащуюся в файле Smoking. База данных содержит информацию о том, были ли введены по месту работы индивида запреты на курение на рабочем месте, является ли тот или иной индивид курильщиком, а также ряд других индивидуальных характеристик.

Данные можно загрузить, используя встроенные набор данных пакета [AER](https://rdrr.io/cran/AER/) (используйте набор данных SmokeBan). Описание данных доступно по [ссылке](https://rdrr.io/cran/AER/man/SmokeBan.html).

**Разведывательный анализ данных**

Проведите разведывательный анализ данных:

1. Постройте таблицу с описательными статистиками основных переменных.
2. Постройте графики, взаимосвязь между возрастом и курением.
3. Постройте таблицу сопряженности между курением и его запретом на рабочем месте.
4. При необходимости проведите преобразования переменных.

**Регрессионный анализ**

Проведите регрессионный анализ, используя модели probit и logit:

1. Оцените коэффициенты в регрессии smoker на ban. Для оценки рекомендуется использовать функцию glm() со следующими аргументами: glm(..., data = ..., family = binomial(link = "..."), x = TRUE).

* Проинтерпретируйте оцененные коэффициенты и их значимость (не забудьте посчитать робастные стандартные ошибки).
* Каков эффект запрета на курение на рабочем месте на вероятность курения для сотрудников?

1. Добавьте в модель контрольные переменные и оцените коэффициенты в новой модели.

* Сравните полученные оценки эффекта запретов на курение с ответом, полученным в предыдущем пункте. Предложите причину, основанную на сущности этой регрессии, которая объясняла бы изменение в предполагаемом эффекте запрета на курение между этими пунктами.
* Проинтерпретируйте оцененные коэффициенты и их значимость (не забудьте посчитать робастные стандартные ошибки).

1. Проанализируйте предельные эффекты для модели с контрольными переменными.

* Каков эффект запрета на курение на рабочем месте на вероятность курения для сотрудников, если зафиксировать все переменные на каком-то уровне? Например, рассмотрите индивида мужского пола среднего возраста, который закончил колледж и не является ни афроамериканцем, ни латиноамеринанцем.
* Рассчитайте и покажите на графике, как отличается эффект запрета на курение для людей разного пола, если остальные переменные зафиксированы.
* Рассчитайте и покажите на графике, как отличается эффект запрета на курение для людей разного возраста, если остальные переменные зафиксированы.

1. (бонус) Рассчитайте оценки отношения шансов и доверительные интервалы к ним для logit модели c контрольными переменными.
2. **Эмпирический пример**

**Вопрос и данные**

Файл [default.csv](https://vpyrlik.github.io/TEACHING/ecox22/datasets/default.csv) содержит финансовую и демографическую информацию о 850 бывших и потенциальных клиентах. Первые 700 наблюдений — это клиенты, которые ранее получали кредиты. Последние 150 наблюдений представляют собой потенциальных клиентов, которых банку нужно классифицировать как хорошие или плохие риски кредитования.



**Задание**

1. Обсудите, какие переменные стоит включить в предсказательную модель.
2. Не забудьте провести разведывательный анализ!
3. Оцените logit-модель и probit-модель вероятности дефолта. Рассмотрите различные спецификации моделей.
4. Рассчитайте предельный эффект от изменения отношения объема задолженности к доходу респондента (DTI) на вероятность дефолта (используйте функции из пакета margins).
5. Проверьте гипотезу о совместной значимости переменных ADDRESS и EMPLOY (используйте функцию linearHypothesis() из пакета car).
6. Рассчитайте предсказанные вероятности для клиентов банка. Предскажите дефолты. Оцените предсказательную силу модели, используя показатели чувствительности и специфичности. Постройте ROC-кривую. Проинтерпретируйте результаты.